

WEB ASOSIDA REAL VAQTLI SO'ROVLARNI QAYTA CHAQARISH UCHUN TEST MA'LUMOTLAR TO'PLAMINI TAYYORLASH VOSITASI

Qudratov Shahzod

Buxoro davlat pedagogika instituti mustaqil izlanuvchisi

e-mail: shahzod100997@gmail.com

Annotatsiya: Ushbu tadqiqotning maqsadi - axborot izlash sohasida yozma hujjatlar to'plamiga asoslangan holda tayyorlangan test ma'lumotlar to'plamlarida, so'rovlarga mos keladigan hujjatlarning relevant (aloqador) yoki norelevant (aloqador emas) deb baholanishini amalga oshiradigan web asosidagi vositani ishlab chiqishdan iborat. Real vaqtlı so'rovni qayta chaqirish uchun tuziladigan test ma'lumotlar to'plami - muayyan hajmdagi hujjatlar korpusi, so'rovlar to'plami hamda ushbu so'rovlarga mos keladigan aloqador hujjatlardan iborat bo'ladi. Xususan, hujjatlarning so'rov maqsadlariga mos yoki mos emasligini baholash jarayoni murakkab va vaqt talab etuvchi bo'lib, bu amaliyot raqamli muhitda amalga oshirilishi zarur. Chunki juda ko'p sonli hujjatlarni aloqadorlik nuqtai nazaridan baholovchi foydalanuvchilar tomonidan ko'rib chiqilishi talab etiladi.

Kalit so'zlar: Test ma'lumotlar to'plami, so'rovni qayta chaqirish, TREC.

Аннотация: Цель данного исследования — разработка веб-ориентированного инструмента, предназначенного для оценки релевантности (соответствия) или нерелевантности (несоответствия) документов в тестовых наборах данных, сформированных на основе коллекций текстовых документов в области информационного поиска. Тестовый набор данных, создаваемый для осуществления поиска по запросу в режиме реального времени, включает в себя корпус документов определённого объёма, набор поисковых запросов и соответствующие этим запросам релевантные документы. В частности, процесс оценки того, соответствуют ли документы цели запроса или нет, является сложным и трудоёмким, и его необходимо выполнять в цифровой среде. Это связано с необходимостью просмотра большого количества документов пользователями, оценивающими их с точки зрения релевантности.

Ключевые слова: Тестовый набор данных, повторный вызов запроса, TREC.

Abstract: The aim of this research is to develop a web-based tool that evaluates the relevance or irrelevance of documents in test datasets created based on collections of written documents in the field of information retrieval. A test dataset designed for real-time query retrieval consists of a document corpus of a certain size, a set of queries, and documents relevant to those queries. Specifically, the process of assessing whether documents match the intent of the queries is complex and time-consuming, and must be

carried out in a digital environment. This is due to the need for users to review a large number of documents to evaluate their relevance.

Keywords: *Test dataset, query retrieval, TREC.*

Kirish

Axborot va kommunikatsiya texnologiyalari jadal rivojlanayotgan bugungi dunyoda insonlarning ehtiyojlari hamda bu ehtiyojlarga topilayotgan yechimlar ham turlicha shakllanmoqda. Deyarli har qanday sohada zarur bo'lgan ma'lumotni internetdagi turli saytlar yoki bloglar orqali topish mumkin. Axborot izlash (AI) vositalari, ayniqsa web-asosidagi qidiruv tizimlari, eng mashhur va eng ko'p foydalaniladigan ilovalar sirasiga kiradi. Yozma hujjatlar ustida axborot izlash jarayoni foydalanuvchining tizimga biror so'rov (query) kiritishi bilan boshlanadi. So'ng, AI tizimi bu so'rov orqali ifodalangan ehtiyojga mos bo'lgan hujjatlarni aloqadorlik (relevance) darajasiga qarab tartiblab, foydalanuvchiga taqdim etadi [1].

Tizimning qayta chaqirish (retrieval) samaradorligi esa taqdim etilgan hujjatlarning foydalanuvchi ehtiyojini qay darajada qondira olganiga bog'liq bo'ladi. AI tizimlarining samaradorligini baholash uchun adabiyotlarda qator o'lchov mezonlari mavjud. Biroq, o'lchov mezonlaridan ham muhimroq masala - bu axborot izlash tizimini qanday test qilishdir. Mazkur yo'nalishda Cranfield tajribalari orqali ishlab chiqilgan AI samaradorligini baholash modeli asosida keyinchalik TREC (Text Retrieval Conference) kabi yirik miqyosdagi tizimlarni baholash ishlari amalga oshirilgan [2].

Ushbu tadqiqotda TREC test derlemlarida hujjatlarning aloqadorligini aniqlash (ya'ni alaka yargisi) jarayoni web asosidagi interaktiv vosita shaklida ishlab chiqilgan. Keyingi bo'limlarda bu jarayonning bosqichlari, vositaning tuzilishi, dizayni va foydalanish usuli batafsil bayon etiladi. Shuningdek, turkcha yozma hujjatlar asosida olib boriladigan axborot izlash tadqiqotlarini rivojlantirish uchun ishonchli va keng qamrovli test ma'lumotlar to'plamlarini yaratish zarur [3].

Motivatsiya

Axborot izlash (ing. *Information Retrieval*, qisqacha AI) - bu fan sohasi foydalanuvchining ma'lumot ehtiyojini qondiradigan, ya'ni aloqador bo'lgan hujjatlarga tez va samarali tarzda erishishni ta'minlash bilan shug'ullanadi. Foydalanuvchining axborot ehtiyoji har doim qat'iy tuzilgan shaklda bo'lmasligi mumkin; ehtiyoj oldindan belgilangan tuzilishga mos ravishda yoki so'rov (query) yoki mavzu sarlavhasi ko'rinishida ifodalangan bo'lishi mumkin [4].

Masalan, so'rovlar istalgan tabiiy tilda yoki mantiqiy ifodalar yordamida shakllantirilishi mumkin. AI sohasidagi dastlabki izlanishlar tabiiy tilda yozilgan matnlar to'plamida matnni qayta chaqirish (text retrieval) asosida olib borilgan. Biroq bugungi kunda ovoz, musiqa, rasm va video kabi ko'p turdagi media hujjatlar bo'yicha ham

shunga o'xshash nazariy asoslar yordamida axborot izlash muammolariga yechim topishga harakat qilinmoqd. Axborot izlash jarayoni amaliyotda qayta chaqirish vazifasi sifatida namoyon bo'ladi.

Ushbu tadqiqotning asosiy maqsadi

Ushbu tadqiqotda maqsad - Axborot izlash (AI) sohasida so'rovlar (instant query) orqali olib boriladigan testlar uchun ma'lumotlar bazasi va amaliy jarayonlarni boshqaruvchi web-ilova ishlab chiqishdan iborat. AI sohasidagi testlar uchun zarur bo'lgan asosiy komponentlar quyidagilardan iborat:

- Vazifaga mos hujjatlar to'plami (korpus),
- So'rovlar to'plami,
- Aloqador hujjatlar to'plami,
- Va baholash mezonlari.

Bu komponentlarning bir-biridan farqlanishi yangi test to'plamlarini yaratish va baholash jarayonini murakkablashtiradi. Ayniqsa, so'rovga aloqador bo'lgan hujjatlarni belgilash (alaka yargisi) ishini raqamli muhitda tez va aniq tarzda amalga oshirish uchun maxsus ilovaga ehtiyoj bor. Sababi, ko'plab hujjatlarni hakam sifatida ishtirok etuvchi foydalanuvchilar tekshirishi kerak bo'ladi. Ana shu muammoni hal etish maqsadida, ushbu tadqiqotda ishlab chiqilgan web-asosli dasturiy vosita bu jarayonni ancha soddalashtiradi va tezlashtiradi [5].

Qaytarish sinov to'plamini tayyorlash jarayoni

TREC (Text REtrieval Conference) - bu axborot izlash tizimlarining samaradorligini baholashga qaratilgan xalqaro tadqiqot tashabbusidir. TREC tomonidan ishlab chiqilgan metodologiyalar axborot izlash sohasida test ma'lumotlar to'plamlarini yaratish va baholash bo'yicha standart hisoblanadi. Instant query asosida axborot qayta chaqirish test to'plamini tayyorlash ushbu metodologiyaga asoslanadi va bir nechta asosiy bosqichlardan iborat [6].

1. Hujjat korpusini shakllantirish

Tajriba uchun mos hujjatlar to'plami (korpus) aniqlanadi. Bu to'plam turli mavzularga oid matnli hujjatlarni o'z ichiga oladi. TRECda bu kabi korpuslar oldindan belgilangan bo'lsa-da, tadqiqotda maxsus lokal hujjatlar to'plami, masalan, yangilik maqolalari, blog yozuvlari yoki ilmiy maqolalar tanlanishi mumkin.

2. So'rovlar (query) bazasini yaratish

Foydalanuvchilarning axborot ehtiyojlarini aks ettiradigan so'rovlar ro'yxati tuziladi. Har bir so'rov oddiy matn ko'rinishida bo'lib, foydalanuvchi tomonidan tizimga kiritilishi mumkin bo'lgan savol yoki izlanish iborasini ifodalaydi.

3. Aloqadorlik (relevance judgment)

So'rovga nisbatan qaysi hujjatlar aloqador (relevant) va qaysilari aloqador emas (non-relevant) ekanini aniqlash jarayonidir. Bu bosqichda tajribali baholovchilar tomonidan

har bir so'rov uchun bir nechta hujjat ko'rib chiqiladi va ular aloqadorlik darajasiga qarab belgilab chiqiladi.

4. Web-asosli yarg'i platformasi

Mazkur tadqiqotda TREC modeliga asoslangan holda ishlab chiqilgan web-ilova orqali baholash jarayoni raqamli tarzda amalga oshirilgan. Bu vosita hakamga har bir so'rov bo'yicha tegishli hujjatlarni ko'rsatadi va ularning aloqadorligini interaktiv tarzda belgilash imkonini beradi. Bu jarayon baholashni soddalashtiradi, tezlashtiradi va natijalarning aniqligini oshiradi.

5. Baholash mezonlari asosida tizim samaradorligini aniqlash

Test to'plami tayyor bo'lgach, axborot izlash tizimi shu asosda sinovdan o'tkaziladi. Precision, Recall, F1-score, MAP (Mean Average Precision) kabi TRECda qo'llaniladigan standart metrikalar asosida tizimning natijalari baholanadi.

Ushbu tadqiqot doirasida ishlab chiqilgan web-asosli hujjat yarg'ilash ilovasi zamonaviy dasturlash texnologiyalari asosida yaratilgan bo'lib, foydalanuvchi uchun qulay interfeysga va barqaror ishlash tizimiga ega. Ilovaning server (backend) qismi Python dasturlash tili va Flask freymvorki yordamida ishlab chiqilgan. Flask yengil va moslashuvchan web freymvork bo'lib, ma'lumotlarni boshqarish, foydalanuvchi so'rovlari va yarg'ilarni qayd etish kabi funksiyalarni bajaradi. Ma'lumotlarni saqlash uchun SQLite bazasi tanlangan bo'lib, SQLAlchemy kutubxonasi orqali Flask bilan integratsiya qilingan. Foydalanuvchi interfeysi (frontend) HTML5, CSS3 va JavaScript texnologiyalariga asoslangan. Dizaynni soddalashtirish va mobil moslashuvchanligini ta'minlash uchun Bootstrap freymvorkidan foydalanilgan. Belgilash, tasdiqlash, yarg'i berish kabi funksiyalar JavaScript orqali dinamik tarzda boshqariladi. Loyihaning barcha kodlari lokal muhitda sinovdan o'tkazilgan bo'lib, kerakli hollarda Git versiya nazorati orqali boshqarilgan. Ushbu texnologiyalar kombinatsiyasi samarali, interaktiv va kengaytiriladigan web-ilovani yaratish imkonini berdi. Platforma orqali foydalanuvchilar hujjatlarni ko'rib chiqib, ularning so'rovga aloqadorligini belgilashlari va shu orqali test to'plamlarini shakllantirishda faol ishtirok etishlari mumkin [7].

Ushbu web-ilovaning dasturiy ta'minot dizayni modulga asoslangan arxitektura tamoyillariga tayangan holda ishlab chiqilgan. Ilovaning asosiy funkcionalligi foydalanuvchilar tomonidan kiritilgan so'rovlarga nisbatan hujjatlarning aloqadorlik darajasini belgilash va ushbu baholar asosida test to'plamlarini shakllantirishdan iborat. Dastur uch asosiy moduldan tashkil topgan: foydalanuvchi interfeysi (frontend), server logikasi (backend) va ma'lumotlar bazasi. Foydalanuvchi interfeysi HTML, CSS va JavaScript yordamida ishlab chiqilgan bo'lib, intuitiv va sodda navigatsiyani ta'minlaydi. Backend qismi esa Python dasturlash tili va Flask freymvorkidan foydalanib yaratilgan. Ushbu qism so'rovlarni qabul qiladi, tegishli ma'lumotlarni bazadan olib, foydalanuvchiga qayta taqdim etadi. Ma'lumotlar bazasi sifatida yengil va portativ

SQLite tanlangan bo'lib, u test hujjatlari, foydalanuvchi yarg'ilari va so'rovlar haqidagi ma'lumotlarni saqlash uchun ishlatiladi. Arxitektura RESTful API tamoyillariga asoslangan bo'lib, foydalanuvchi bilan server o'rtasida samarali va xavfsiz ma'lumot almashinuvi amalga oshiriladi. Dizayn davomida tizim xavfsizligi, foydalanuvchi maxfiyligi va ma'lumot yaxlitligi ustuvor yo'nalish sifatida qaraldi. Ilova kengaytiriladigan va yangilanadigan tarzda ishlab chiqilgan bo'lib, kelgusida boshqa tillar yoki ma'lumotlar formati bilan integratsiyalash imkonini ham ta'minlaydi.

Xulosa

Mazkur tadqiqotda aniq va real so'rovlar asosida test ma'lumotlar to'plamlarini yaratish imkonini beruvchi web-vositaning ishlab chiqilishi va qo'llanilishi yoritildi. An'anaviy yondashuvlarda test to'plamlarining yaratilishi ko'p vaqt talab qiladigan, murakkab va ko'p resurs talab etuvchi jarayon bo'lib kelgan. Ayniqsa, so'rovlarga nisbatan aloqadorlik bahosining qo'lda belgilanishi subyektiv yondashuvga sabab bo'lishi mumkin edi. Ushbu web-ilovaning ishlab chiqilishi orqali bu jarayon avtomatlashtirilib, foydalanuvchilarga tez, qulay va tizimli ravishda baholash imkoniyati yaratildi.

Tadqiqot davomida TREC (Text Retrieval Conference) modeli asosida sinov o'tkazish va aloqadorlikni belgilash tamoyillari asos qilib olindi. Ilova orqali foydalanuvchilar bir vaqtning o'zida ko'plab hujjatlarni ko'rib chiqib, ularning so'rov bilan bog'liqligini belgilangach, bu baholashlar avtomatik tarzda test to'plamiga qo'shiladi. Bu yondashuv nafaqat vaqtini tejaydi, balki test natijalarining ishonchliligini ham oshiradi.

Xulosa sifatida aytish mumkinki, ishlab chiqilgan web-vosita nafaqat tizimlari uchun sinov ma'lumotlarini tayyorlashni soddalashtiradi, balki tadqiqotchilar va talabalarning ilmiy ishlanmalarida sifatli va aniqligi ta'minlangan test to'plamlaridan foydalanishiga ham xizmat qiladi. Kelgusidagi ishlarda ushbu tizim boshqa tillar, yangi formatdagi hujjatlar va ko'p foydalanuvchili interaktivlik bilan kengaytirilishi mumkin. Bu esa uni nafaqat mahalliy, balki xalqaro miqyosda qo'llanishi mumkin bo'lgan foydali vositaga aylantiradi.

Foydalanilgan adabiyotlar

1. Salton, G., & McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc.
2. Harman, D. (1995). *The TREC Test Collection*. In *Information Processing & Management*, 31(3), 327–340.
3. Buckley, C., & Voorhees, E. M. (2000). *Evaluating Evaluation Measure Stability*. In *Proceedings of the 23rd Annual International ACM SIGIR Conference*.

4. Croft, W. B., Metzler, D., & Strohman, T. (2010). *Search Engines: Information Retrieval in Practice*. Addison-Wesley.
5. Robertson, S. E., & Zaragoza, H. (2009). *The Probabilistic Relevance Framework: BM25 and Beyond*. Foundations and Trends in Information Retrieval, 3(4), 333–389.
6. Brants, T., & Franz, A. (2006). *Web 1T 5-gram Version 1*. Linguistic Data Consortium.
7. Allen, I. E., & Seaman, J. (2011). *Going the Distance: Online Education in the United States*. Babson Survey Research Group.

